

SIGNAL PROCESSING SYSTEM

The present invention relates to a signal processing method and apparatus. The invention is particularly relevant to a statistical analysis of signals output by a plurality of sensors in response to signals generated by a plurality of sources. The invention may be used in speech applications and in other applications to process the received signals in order to separate the signals generated by the plurality of sources. The invention can also be used to identify the number of sources that are present.

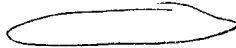
There exists a need to be able to process signals output by a plurality of sensors in response to signals generated by a plurality of sources. The sources may, for example, be different users speaking and the sensors may be microphones. Current techniques employ arrays of microphones and an adaptive beam forming technique in order to isolate the speech from one of the speakers. This kind of beam forming system suffers from a number of problems. Firstly, it can only isolate signals from sources that are spatially distinct. It also does not work if the sources are relatively close together since the "beam" which it uses has a finite resolution. It is also necessary to know the directions from which the

signals of interest will arrive and also the spacing between the sensors in the sensor array. Further, if N sensors are available, then only $N - 1$ "nulls" can be created within the sensing zone.

5

An aim of the present invention is to provide an alternative technique for processing the signals output from a plurality of sensors in response to signals received from a plurality of sources.

10



15

According to one aspect, the present invention provides a signal processing apparatus comprising: one or more receivers for receiving a set of signal values representative of signals generated by a plurality of signal sources; a memory for storing a probability density function for parameters of a respective signal model, each of which is assumed to have generated a respective one of the signals represented by the received signal values; means for applying the received signal values to the probability density function; means for processing the probability density function with those values applied to derive samples of parameter values from the probability density function; and means for analysing some of the derived samples to determine parameter values that are representative of the signals generated by at

20

25

least one of the sources.



Exemplary embodiments of the present invention will now be described with reference to the accompanying drawings in which:

Figure 1 is a schematic view of a computer which may be programmed to operate in accordance with an embodiment of the present invention;

Figure 2 is a block diagram illustrating the principal components of a speech recognition system;

Figure 3 is a block diagram representing a model employed by a statistical analysis unit which forms part of the speech recognition system shown in Figure 2;

Figure 4 is a flow chart illustrating the processing steps performed by a model order selection unit forming part of the statistical analysis unit shown in Figure 2;

Figure 5 is a flow chart illustrating the main processing steps employed by a Simulation Smoother which forms part of the statistical analysis unit shown in Figure 2;

Figure 6 is a block diagram illustrating the main processing components of the statistical analysis unit shown in Figure 2;

5 Figure 7 is a memory map illustrating the data that is stored in a memory which forms part of the statistical analysis unit shown in Figure 2;

10 Figure 8 is a flow chart illustrating the main processing steps performed by the statistical analysis unit shown in Figure 6;

15 Figure 9a is a histogram for a model order of an autoregressive filter model which forms part of the model shown in Figure 3;

Figure 9b is a histogram for the variance of process noise modelled by the model shown in Figure 3;

20 Figure 9c is a histogram for a third coefficient of the AR filter model;

25 Figure 10 is a block diagram illustrating the principal components of a speech recognition system embodying the present invention;

Figure 11 is a block diagram representing a model employed by a statistical analysis unit which forms part of the speech recognition system shown in Figure 10;

5 Figure 12 is block diagram illustrating the principal components of a speech recognition system embodying the present invention;

10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
222

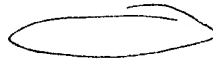
Figure 16 is a block diagram illustrating the main components of an alternative speech recognition system in which data output by the statistical analysis unit is used to detect the beginning and end of speech within the input signal;

Figure 17 is a schematic block diagram illustrating the principal components of a speaker verification system;

Figure 18 is a schematic block diagram illustrating the principal components of an acoustic classification system;

Figure 19 is a schematic block diagram illustrating the principal components of a speech encoding and transmission; and

Figure 20 is a block diagram illustrating the principal components of a data file annotation system which uses the statistical analysis unit shown in Figure 6 to provide quality of speech data for an associated annotation.



Embodiments of the present invention can be implemented on computer hardware, but the embodiment to be described

is implemented in software which is run in conjunction with processing hardware such as a personal computer, workstation, photocopier, facsimile machine or the like.

5 Figure 1 is a personal computer (PC) 1 which may be programmed to operate an embodiment of the present invention. A keyboard 3, a pointing device 5, two microphones 7-1 and 7-2 and a telephone-line 9 are connected to the PC 1 via an interface 11. A keyboard 3 and pointing device 5 allow the system to be controlled by a user. The microphones 7 convert the acoustic speech signal of one or more users into equivalent electrical signals and supplies them to the PC 1 for processing. An internal modem and speech receiving circuit (not shown) may be connected to the telephone line 9 so that the PC 1 can communicate with, for example, a remote computer or with a remote user.

20 The program instructions which make the PC 1 operate in accordance with the present invention may be supplied for use with an existing PC 1 on, for example, a storage device such as a magnetic disc 13, or by downloading the software from the Internet (not shown) via the internal modem and telephone line 9.

The operation of a speech recognition system which receives signals output from multiple microphones in response to speech signals generated from a plurality of speakers will be described. However, in order to facilitate the understanding of the operation of such a recognition system, a speech recognition system which performs a similar analysis of the signals output from the microphone for the case of a single speaker and single microphone will be described first with reference to Figure 2 to 9.

SINGLE SPEAKER SINGLE MICROPHONE

As shown in Figure 2, electrical signals representative of the input speech from the microphone 7 are input to a filter 15 which removes unwanted frequencies (in this embodiment frequencies above 8 kHz) within the input signal. The filtered signal is then sampled (at a rate of 16 kHz) and digitised by the analogue to digital converter 17 and the digitised speech samples are then stored in a buffer 19. Sequential blocks (or frames) of speech samples are then passed from the buffer 19 to a statistical analysis unit 21 which performs a statistical analysis of each frame of speech samples in sequence to determine, amongst other things, a set of auto regressive (AR) coefficients representative of the speech within the

frame. In this embodiment, the AR coefficients output by the statistical analysis unit 21 are then input, via a coefficient converter 23 to a cepstral based speech recognition unit 25. In this embodiment, therefore, the coefficient converter 23 converts the AR coefficients output by the analysis unit 21 into cepstral coefficients. This can be achieved using the conversion technique described in, for example, "Fundamentals of Speech Recognition" by Rabiner and Juang at pages 115 and 116. The speech recognition unit 25 then compares the cepstral coefficients for successive frames of speech with a set of stored speech models 27, which may be template based or Hidden Markov Model based, to generate a recognition result.

Statistical Analysis Unit - Theory and Overview

As mentioned above, the statistical analysis unit 21 analyses the speech within successive frames of the input speech signal. In most speech processing systems, the frames are overlapping. However, in this embodiment, the frames of speech are non-overlapping and have a duration of 20ms which, with the 16kHz sampling rate of the analogue to digital converter 17, results in a frame size of 320 samples.

In order to perform the statistical analysis on each of the frames, the analysis unit 21 assumes that there is an underlying process which generated each sample within the frame. The model of this process used in this embodiment is shown in Figure 3. As shown, the process is modelled by a speech source 31 which generates, at time $t = n$, a raw speech sample $s(n)$. Since there are physical constraints on the movement of the speech articulators, there is some correlation between neighbouring speech samples. Therefore, in this embodiment, the speech source 31 is modelled by an auto regressive (AR) process. In other words, the statistical analysis unit 21 assumes that a current raw speech sample ($s(n)$) can be determined from a linear weighted combination of the most recent previous raw speech samples, i.e.:

$$s(n) = a_1s(n-1) + a_2s(n-2) + \dots + a_k s(n-k) + e(n) \quad (1)$$

where a_1, a_2, \dots, a_k are the AR filter coefficients representing the amount of correlation between the speech samples; k is the AR filter model order; and $e(n)$ represents random process noise which is involved in the generation of the raw speech samples. As those skilled in the art of speech processing will appreciate, these AR filter coefficients are the same coefficients that the

linear prediction (LP) analysis estimates albeit using a different processing technique.

As shown in Figure 3, the raw speech samples $s(n)$ generated by the speech source are input to a channel 33 which models the acoustic environment between the speech source 31 and the output of the analogue to digital converter 17. Ideally, the channel 33 should simply attenuate the speech as it travels from the source 31 to the microphone 7. However, due to reverberation and other distortive effects, the signal $(y(n))$ output by the analogue to digital converter 17 will depend not only on the current raw speech sample $(s(n))$ but it will also depend upon previous raw speech samples. Therefore, in this embodiment, the statistical analysis unit 21 models the channel 33 by a moving average (MA) filter, i.e.:

$$y(n) = h_0 s(n) + h_1 s(n-1) + h_2 s(n-2) + \dots + h_r s(n-r) + \varepsilon(n) \quad (2)$$

where $y(n)$ represents the signal sample output by the analogue to digital converter 17 at time $t = n$; $h_0, h_1, h_2, \dots, h_r$ are the channel filter coefficients representing the amount of distortion within the channel 33; r is the channel filter model order; and $\varepsilon(n)$ represents a random additive measurement noise component.

For the current frame of speech being processed, the filter coefficients for both the speech source and the channel are assumed to be constant but unknown. Therefore, considering all N samples (where $N = 320$) in the current frame being processed gives:

$$\begin{aligned}
 s(n) &= a_1 s(n-1) + a_2 s(n-2) + \dots + a_k s(n-k) + e(n) \\
 s(n-1) &= a_1 s(n-2) + a_2 s(n-3) + \dots + a_k s(n-k-1) + e(n-1) \\
 &\vdots \\
 s(n-N+1) &= a_1 s(n-N) + a_2 s(n-N-1) + \dots + a_k s(n-k-N+1) + e(n-N+1)
 \end{aligned} \tag{3}$$

which can be written in vector form as:

$$\underline{s}(n) = S \underline{a} + \underline{e}(n) \tag{4}$$

where

$$S = \begin{bmatrix} s(n-1) & s(n-2) & s(n-3) & \dots & s(n-k) \\ s(n-2) & s(n-3) & s(n-4) & \dots & s(n-k-1) \\ s(n-3) & s(n-4) & s(n-5) & \dots & s(n-k-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s(n-N) & s(n-N-1) & s(n-N-2) & \dots & s(n-k-N+1) \end{bmatrix}_{N \times k}$$

and

$$\underline{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_k \end{bmatrix}_{k \times 1} \quad \underline{s}(n) = \begin{bmatrix} s(n) \\ s(n-1) \\ s(n-2) \\ \vdots \\ s(n-N+1) \end{bmatrix}_{N \times 1} \quad \underline{e}(n) = \begin{bmatrix} e(n) \\ e(n-1) \\ e(n-2) \\ \vdots \\ e(n-N+1) \end{bmatrix}_{N \times 1}$$

As will be apparent from the following discussion, it is also convenient to rewrite equation (3) in terms of the random error component (often referred to as the residual) $e(n)$. This gives:

$$\begin{aligned}
 e(n) &= s(n) - a_1 s(n-1) - a_2 s(n-2) - \dots - a_k s(n-k) \\
 e(n-1) &= s(n-1) - a_1 s(n-2) - a_2 s(n-3) - \dots - a_k s(n-k-1) \\
 &\vdots \\
 e(n-N+1) &= s(n-N+1) - a_1 s(n-N) - a_2 s(n-N-1) - \dots - a_k s(n-k-N+1)
 \end{aligned} \tag{5}$$

which can be written in vector notation as:

$$\underline{e}(n) = \underline{A} \underline{s}(n) \tag{6}$$

where

$$\underline{A} = \begin{bmatrix} 1 & -a_1 & -a_2 & -a_3 & \dots & -a_k & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & -a_1 & -a_2 & \dots & -a_{k-1} & -a_k & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & -a_1 & \dots & -a_{k-2} & -a_{k-1} & -a_k & 0 & \dots & 0 \\ \vdots & & & & & & & & & & \\ \vdots & & & & & & & & & & \\ \vdots & & & & & & & & & & \\ 0 & & & & & & & & & & 1 \end{bmatrix}_{N \times N}$$

Similarly, considering the channel model defined by equation (2), with $h_0 = 1$ (since this provides a more stable solution), gives:

$$\begin{aligned}
 q(n) &= h_1 s(n-1) + h_2 s(n-2) + \dots + h_r s(n-r) + \varepsilon(n) \\
 q(n-1) &= h_1 s(n-2) + h_2 s(n-3) + \dots + h_r s(n-r-1) + \varepsilon(n-1) \\
 &\vdots \\
 q(n-N+1) &= h_1 s(n-N) + h_2 s(n-N-1) + \dots + h_r s(n-r-N+1) + \varepsilon(n-N+1)
 \end{aligned} \tag{7}$$

(where $q(n) = y(n) - s(n)$) which can be written in vector form as:

$$q(n) = Y \cdot h + \varepsilon(n) \quad (8)$$

where

$$Y = \begin{bmatrix} s(n-1) & s(n-2) & s(n-3) & \dots & s(n-r) \\ s(n-2) & s(n-3) & s(n-4) & \dots & s(n-r-1) \\ s(n-3) & s(n-4) & s(n-5) & \dots & s(n-r-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s(n-N) & s(n-N-1) & s(n-N-2) & \dots & s(n-r-N+1) \end{bmatrix}_{N \times r}$$

and

$$h = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ \vdots \\ h_r \end{bmatrix}_{r \times 1} \quad q(n) = \begin{bmatrix} q(n) \\ q(n-1) \\ q(n-2) \\ \vdots \\ q(n-N+1) \end{bmatrix}_{N \times 1} \quad \varepsilon(n) = \begin{bmatrix} \varepsilon(n) \\ \varepsilon(n-1) \\ \varepsilon(n-2) \\ \vdots \\ \varepsilon(n-N+1) \end{bmatrix}_{N \times 1}$$

In this embodiment, the analysis unit 21 aims to determine, amongst other things, values for the AR filter coefficients (\underline{a}) which best represent the observed signal samples ($y(n)$) in the current frame. It does this by determining the AR filter coefficients (\underline{a}) that maximise the joint probability density function of the speech model, channel model, raw speech samples and the noise statistics given the observed signal samples output from the analogue to digital converter 17, i.e. by determining:

$$\max_a \left\{ p(a, k, h, r, \sigma_e^2, \sigma_v^2, s(n) | y(n)) \right\} \quad (9)$$

where σ_e^2 and σ_v^2 represent the process and measurement noise statistics respectively. As those skilled in the art will appreciate, this function defines the probability that a particular speech model, channel model, raw speech samples and noise statistics generated the observed frame of speech samples ($y(n)$) from the analogue to digital converter. To do this, the statistical analysis unit 21 must determine what this function looks like. This problem can be simplified by rearranging this probability density function using Bayes law to give:

$$\frac{p(y(n) | s(n), h, r, \sigma_e^2) p(s(n) | a, k, \sigma_v^2) p(a | k) p(h | r) p(\sigma_e^2) p(\sigma_v^2) p(k) p(r)}{p(y(n))} \quad (10)$$

As those skilled in the art will appreciate, the denominator of equation (10) can be ignored since the probability of the signals from the analogue to digital converter is constant for all choices of model. Therefore, the AR filter coefficients that maximise the function defined by equation (9) will also maximise the numerator of equation (10).

Each of the terms on the numerator of equation (10) will now be considered in turn.

$$p(\underline{s}(n) | \underline{a}, k, \sigma_e^2)$$

5 This term represents the joint probability density function for generating the vector of raw speech samples ($\underline{s}(n)$) during a frame, given the AR filter coefficients (\underline{a}), the AR filter model order (k) and the process noise statistics (σ_e^2). From equation (6) above, this joint
10 probability density function for the raw speech samples can be determined from the joint probability density function for the process noise. In particular $p(\underline{s}(n) | \underline{a}, k, \sigma_e^2)$ is given by:

$$15 \quad p(\underline{s}(n) | \underline{a}, k, \sigma_e^2) = p(\underline{e}(n)) \left| \frac{\delta \underline{e}(n)}{\delta \underline{s}(n)} \right|_{\underline{e}(n) = \underline{s}(n) - S\underline{a}} \quad (11)$$

where $p(\underline{e}(n))$ is the joint probability density function for the process noise during a frame of the input speech and the second term on the right-hand side is known as
20 the Jacobean of the transformation. In this case, the Jacobean is unity because of the triangular form of the matrix \ddot{A} (see equations (6) above).

In this embodiment, the statistical analysis unit 21
25 assumes that the process noise associated with the speech

source 31 is Gaussian having zero mean and some unknown variance σ_e^2 . The statistical analysis unit 21 also assumes that the process noise at one time point is independent of the process noise at another time point. Therefore, the joint probability density function for the process noise during a frame of the input speech (which defines the probability of any given vector of process noise $\underline{e}(n)$ occurring) is given by:

$$p(\underline{e}(n)) = (2\pi\sigma_e^2)^{-\frac{N}{2}} \exp\left[\frac{-\underline{e}(n)^T \underline{e}(n)}{2\sigma_e^2}\right] \quad (12)$$

Therefore, the joint probability density function for a vector of raw speech samples given the AR filter coefficients (\underline{a}), the AR filter model order (k) and the process noise variance (σ_e^2) is given by:

$$p(\underline{s}(n)|\underline{a}, k, \sigma_e^2) = (2\pi\sigma_e^2)^{-\frac{N}{2}} \exp\left[\frac{-1}{2\sigma_e^2} \left(\underline{s}(n)^T \underline{s}(n) - 2\underline{a}^T S \underline{s}(n) + \underline{a}^T S^T S \underline{a}\right)\right] \quad (13)$$

$$p(\underline{y}(n)|\underline{s}(n), \underline{h}, r, \sigma_e^2)$$

This term represents the joint probability density function for generating the vector of speech samples ($\underline{y}(n)$) output from the analogue to digital converter 17, given the vector of raw speech samples ($\underline{s}(n)$), the channel filter coefficients (\underline{h}), the channel filter model order (r) and the measurement noise statistics (σ_e^2).

From equation (8), this joint probability density function can be determined from the joint probability density function for the process noise. In particular, $p(y(n)|\underline{s}(n), \underline{h}, r, \sigma_e^2)$ is given by:

$$p(y(n)|\underline{s}(n), \underline{h}, r, \sigma_e^2) = p(\underline{\varepsilon}(n)) \left| \frac{\delta \underline{\varepsilon}(n)}{\delta y(n)} \right|_{\underline{\varepsilon}(n) = q(n) - Yh} \quad (14)$$

where $p(\underline{\varepsilon}(n))$ is the joint probability density function for the measurement noise during a frame of the input speech and the second term on the right hand side is the Jacobean of the transformation which again has a value of one.

In this embodiment, the statistical analysis unit 21 assumes that the measurement noise is Gaussian having zero mean and some unknown variance σ_e^2 . It also assumes that the measurement noise at one time point is independent of the measurement noise at another time point. Therefore, the joint probability density function for the measurement noise in a frame of the input speech will have the same form as the process noise defined in equation (12). Therefore, the joint probability density function for a vector of speech samples $(y(n))$ output from the analogue to digital converter 17, given the channel filter coefficients (\underline{h}) , the channel filter model

order (r), the measurement noise statistics (σ_e^2) and the raw speech samples ($\underline{s}(n)$) will have the following form:

$$p(\underline{y}(n)|\underline{s}(n), h, r, \sigma_e^2) = (2\pi\sigma_e^2)^{-\frac{N}{2}} \exp \left[\frac{-1}{2\sigma_e^2} \left(\underline{q}(n)^T \underline{q}(n) - 2h^T Y \underline{q}(n) + h^T Y^T Y h \right) \right] \quad (15)$$

As those skilled in the art will appreciate, although this joint probability density function for the vector of speech samples ($\underline{y}(n)$) is in terms of the variable $\underline{q}(n)$, this does not matter since $\underline{q}(n)$ is a function of $\underline{y}(n)$ and $\underline{s}(n)$, and $\underline{s}(n)$ is a given variable (ie known) for this probability density function.

$p(\underline{a}|k)$

This term defines the *prior* probability density function for the AR filter coefficients (\underline{a}) and it allows the statistical analysis unit 21 to introduce knowledge about what values it expects these coefficients will take. In this embodiment, the statistical analysis unit 21 models this prior probability density function by a Gaussian having an unknown variance (σ_a^2) and mean vector ($\underline{\mu}_a$), i.e.:

$$p(\underline{a}|k, \sigma_a^2, \underline{\mu}_a) = (2\pi\sigma_a^2)^{-\frac{N}{2}} \exp \left[\frac{-(\underline{a} - \underline{\mu}_a)^T (\underline{a} - \underline{\mu}_a)}{2\sigma_a^2} \right] \quad (16)$$

By introducing the new variables σ_a^2 and $\underline{\mu}_a$, the prior

density functions ($p(\sigma_a^2)$ and $p(\underline{\mu}_a)$) for these variables must be added to the numerator of equation (10) above. Initially, for the first frame of speech being processed the mean vector ($\underline{\mu}_a$) can be set to zero and for the

5 second and subsequent frames of speech being processed, it can be set to the mean vector obtained during the processing of the previous frame. In this case, $p(\underline{\mu}_a)$ is just a Dirac delta function located at the current value of $\underline{\mu}_a$ and can therefore be ignored.

10 With regard to the prior probability density function for the variance of the AR filter coefficients, the statistical analysis unit 21 could set this equal to some constant to imply that all variances are equally

15 probable. However, this term can be used to introduce knowledge about what the variance of the AR filter coefficients is expected to be. In this embodiment, since variances are always positive, the statistical analysis unit 21 models this variance prior probability

20 density function by an Inverse Gamma function having parameters α_a and β_a , i.e.:

$$p(\sigma_a^2 | \alpha_a, \beta_a) = \frac{(\sigma_a^2)^{-(\alpha_a + 1)}}{\beta_a \Gamma(\alpha_a)} \exp \left[\frac{-1}{\sigma_a^2 \beta_a} \right] \quad (17)$$

25 At the beginning of the speech being processed, the

statistical analysis unit 21 will not have much knowledge about the variance of the AR filter coefficients. Therefore, initially, the statistical analysis unit 21 sets the variance σ_a^2 and the α and β parameters of the Inverse Gamma function to ensure that this probability density function is fairly flat and therefore non-informative. However, after the first frame of speech has been processed, these parameters can be set more accurately during the processing of the next frame of speech by using the parameter values calculated during the processing of the previous frame of speech.

$p(\underline{h}|\underline{r})$

This term represents the *prior* probability density function for the channel model coefficients (\underline{h}) and it allows the statistical analysis unit 21 to introduce knowledge about what values it expects these coefficients to take. As with the prior probability density function for the AR filter coefficients, in this embodiment, this probability density function is modelled by a Gaussian having an unknown variance (σ_h^2) and mean vector ($\underline{\mu}_h$), i.e.:

$$p(\underline{h}|\underline{r}, \sigma_h^2, \underline{\mu}_h) = (2\pi\sigma_h^2)^{-\frac{N}{2}} \exp\left[\frac{-(\underline{h} - \underline{\mu}_h)^T(\underline{h} - \underline{\mu}_h)}{2\sigma_h^2}\right] \quad (18)$$

Again, by introducing these new variables, the prior density functions ($p(\sigma_h)$ and $p(\underline{\mu}_h)$) must be added to the numerator of equation (10). Again, the mean vector can initially be set to zero and after the first frame of speech has been processed and for all subsequent frames of speech being processed, the mean vector can be set to equal the mean vector obtained during the processing of the previous frame. Therefore, $p(\underline{\mu}_h)$ is also just a Dirac delta function located at the current value of $\underline{\mu}_h$ and can be ignored.

With regard to the *prior* probability density function for the variance of the channel filter coefficients, again, in this embodiment, this is modelled by an Inverse Gamma function having parameters α_h and β_h . Again, the variance (σ_h^2) and the α and β parameters of the Inverse Gamma function can be chosen initially so that these densities are non-informative so that they will have little effect on the subsequent processing of the initial frame.

$p(\sigma_e^2)$ and $p(\sigma_c^2)$

These terms are the *prior* probability density functions for the process and measurement noise variances and again, these allow the statistical analysis unit 21 to introduce knowledge about what values it expects these

noise variances will take. As with the other variances, in this embodiment, the statistical analysis unit 21 models these by an Inverse Gamma function having parameters α_e , β_e and α_ε , β_ε respectively. Again, these variances and these Gamma function parameters can be set initially so that they are non-informative and will not appreciably affect the subsequent calculations for the initial frame.

p(k) and p(r)

These terms are the *prior* probability density functions for the AR filter model order (k) and the channel model order (r) respectively. In this embodiment, these are modelled by a uniform distribution up to some maximum order. In this way, there is no prior bias on the number of coefficients in the models except that they can not exceed these predefined maximums. In this embodiment, the maximum AR filter model order (k) is thirty and the maximum channel model order (r) is one hundred and fifty.

Therefore, inserting the relevant equations into the numerator of equation (10) gives the following joint probability density function which is proportional to $p(\underline{a}, k, \underline{h}, r, \sigma_a^2, \sigma_h^2, \sigma_e^2, \sigma_\varepsilon^2, \underline{s}(n) | \underline{y}(n))$:

5

$$\begin{aligned}
& (2\pi\sigma_e^2)^{-\frac{N}{2}} \exp \left[\frac{-1}{2\sigma_e^2} \left(\underline{q}(n)^T \underline{q}(n) - 2\mathbf{h}^T \mathbf{Y} \underline{q}(n) + \mathbf{h}^T \mathbf{Y}^T \mathbf{Y} \mathbf{h} \right) \right] \\
& \times (2\pi\sigma_e^2)^{-\frac{N}{2}} \exp \left[\frac{-1}{2\sigma_e^2} \left(\underline{s}(n)^T \underline{s}(n) - 2\mathbf{a}^T \mathbf{S} \underline{s}(n) + \mathbf{a}^T \mathbf{S}^T \mathbf{S} \mathbf{a} \right) \right] \\
& \times (2\pi\sigma_a^2)^{-\frac{N}{2}} \exp \left[\frac{-(\mathbf{a} - \underline{\mu}_a)^T (\mathbf{a} - \underline{\mu}_a)}{2\sigma_a^2} \right] \times (2\pi\sigma_h^2)^{-\frac{N}{2}} \exp \left[\frac{-(\mathbf{h} - \underline{\mu}_h)^T (\mathbf{h} - \underline{\mu}_h)}{2\sigma_h^2} \right] \\
& \times \frac{(\sigma_a^2)^{-(\alpha_a+1)}}{\beta_a \Gamma(\alpha_a)} \exp \left[\frac{-1}{\sigma_a^2 \beta_a} \right] \times \frac{(\sigma_h^2)^{-(\alpha_h+1)}}{\beta_h \Gamma(\alpha_h)} \exp \left[\frac{-1}{\sigma_h^2 \beta_h} \right] \\
& \times \frac{(\sigma_e^2)^{-(\alpha_e+1)}}{\beta_e \Gamma(\alpha_e)} \exp \left[\frac{-1}{\sigma_e^2 \beta_e} \right] \times \frac{(\sigma_e^2)^{-(\alpha_e+1)}}{\beta_e \Gamma(\alpha_e)} \exp \left[\frac{-1}{\sigma_e^2 \beta_e} \right]
\end{aligned}$$

10

(19)

Gibbs Sampler

15

In order to determine the form of this joint probability density function, the statistical analysis unit 21 "draws samples" from it. In this embodiment, since the joint probability density function to be sampled is a complex multivariate function, a Gibbs sampler is used which breaks down the problem into one of drawing samples from probability density functions of smaller dimensionality. In particular, the Gibbs sampler proceeds by drawing random variates from conditional densities as follows:

20

25

first iteration

$$p(\underline{a}, k | h^0, r^0, \sigma_e^{2^0}, \sigma_\varepsilon^{2^0}, \sigma_a^{2^0}, \sigma_h^{2^0}, \underline{s}(n)^0, y(n)) \rightarrow \underline{a}^1, k^1$$

$$p(h, r | \underline{a}^1, k^1, \sigma_e^{2^0}, \sigma_\varepsilon^{2^0}, \sigma_a^{2^0}, \sigma_h^{2^0}, \underline{s}(n)^0, y(n)) \rightarrow h^1, r^1$$

$$p(\sigma_e^2 | \underline{a}^1, k^1, h^1, r^1, \sigma_e^{2^0}, \sigma_\varepsilon^{2^0}, \sigma_a^{2^0}, \sigma_h^{2^0}, \underline{s}(n)^0, y(n)) \rightarrow \sigma_e^{2^1}$$

$$p(\sigma_h^{2^1} | \underline{a}^1, k^1, h^1, r^1, \sigma_e^{2^1}, \sigma_\varepsilon^{2^1}, \sigma_a^{2^1}, \sigma_h^{2^1}, \underline{s}(n)^0, y(n)) \rightarrow \sigma_h^{2^1}$$

second iteration

$$p(\underline{a}, k | h^1, r^1, \sigma_e^{2^1}, \sigma_\varepsilon^{2^1}, \sigma_h^{2^1}, \underline{s}(n)^1, y(n)) \rightarrow \underline{a}^2, k^2$$

$$p(h, r | \underline{a}^2, k^2, \sigma_e^{2^1}, \sigma_\varepsilon^{2^1}, \sigma_a^{2^1}, \sigma_h^{2^1}, \underline{s}(n)^1, y(n)) \rightarrow h^2, r^2$$

etc.

where $(h^0, r^0, (\sigma_e^2)^0, (\sigma_\varepsilon^2)^0, (\sigma_a^2)^0, (\sigma_h^2)^0, \underline{s}(n)^0)$ are initial values which may be obtained from the results of the statistical analysis of the previous frame of speech, or where there are no previous frames, can be set to appropriate values that will be known to those skilled in the art of speech processing.

As those skilled in the art will appreciate, these conditional densities are obtained by inserting the

current values for the given (or known) variables into the terms of the density function of equation (19). For the conditional density $p(\underline{a}, k | \dots)$ this results in:

$$\begin{aligned}
 p(\underline{a}, k | \dots) \propto \exp \left[\frac{-1}{2\sigma_e^2} \left(\underline{s}(n)^T \underline{s}(n) - 2\underline{a}^T S \underline{s}(n) + \underline{a}^T S^T S \underline{a} \right) \right] \\
 \times \exp \left[\frac{-(\underline{a} - \underline{\mu}_a)^T (\underline{a} - \underline{\mu}_a)}{2\sigma_a^2} \right]
 \end{aligned} \tag{20}$$

which can be simplified to give:

$$p(\underline{a}, k | \dots) \propto \exp \left[\frac{-1}{2} \left(\frac{\underline{s}(n)^T \underline{s}(n)}{\sigma_e^2} + \frac{\underline{\mu}_a^T \underline{\mu}_a}{\sigma_a^2} - 2\underline{a}^T \left[\frac{S \underline{s}(n)}{\sigma_e^2} + \frac{\underline{\mu}_a}{\sigma_a^2} \right] + \underline{a}^T \left[\frac{S^T S}{\sigma_e^2} + \frac{I}{\sigma_a^2} \right] \underline{a} \right) \right] \tag{21}$$

which is in the form of a standard Gaussian distribution having the following covariance matrix:

$$\Sigma_a = \left[\frac{S^T S}{\sigma_e^2} + \frac{I}{\sigma_a^2} \right]^{-1} \tag{22}$$

The mean value of this Gaussian distribution can be determined by differentiating the exponent of equation (21) with respect to \underline{a} and determining the value of \underline{a} which makes the differential of the exponent equal to

zero. This yields a mean value of:

$$\hat{\mu}_a = \left[\frac{S^T S}{\sigma_e^2} + \frac{I}{\sigma_a^2} \right]^{-1} \left[\frac{S^T s(n)}{\sigma_e^2} + \frac{\mu_a}{\sigma_a^2} \right] \quad (23)$$

5 A sample can then be drawn from this standard Gaussian distribution to give \underline{a}^g (where g is the g^{th} iteration of the Gibbs sampler) with the model order (k^g) being determined by a model order selection routine which will be described later. The drawing of a sample from this
10 Gaussian distribution may be done by using a random number generator which generates a vector of random values which are uniformly distributed and then using a transformation of random variables using the covariance matrix and the mean value given in equations (22) and
15 (23) to generate the sample. In this embodiment, however, a random number generator is used which generates random numbers from a Gaussian distribution having zero mean and a variance of one. This simplifies the transformation process to one of a simple scaling
20 using the covariance matrix given in equation (22) and shifting using the mean value given in equation (23). Since the techniques for drawing samples from Gaussian distributions are well known in the art of statistical analysis, a further description of them will not be given
25 here. A more detailed description and explanation can be

found in the book entitled "Numerical Recipes in C", by W. Press et al, Cambridge University Press, 1992 and in particular at chapter 7.

As those skilled in the art will appreciate, however, before a sample can be drawn from this Gaussian distribution, estimates of the raw speech samples must be available so that the matrix S and the vector $\underline{s}(n)$ are known. The way in which these estimates of the raw speech samples are obtained in this embodiment will be described later.

A similar analysis for the conditional density $p(\underline{h}, r | \dots)$ reveals that it also is a standard Gaussian distribution but having a covariance matrix and mean value given by:

$$\Sigma_h = \left[\frac{Y^T Y}{\sigma_e^2} + \frac{I}{\sigma_h^2} \right]^{-1} \quad \hat{\mu}_h = \left[\frac{Y^T Y}{\sigma_e^2} + \frac{I}{\sigma_h^2} \right]^{-1} \left[\frac{Y^T q(n)}{\sigma_e^2} + \frac{\mu_h}{\sigma_h^2} \right] \quad (24)$$

from which a sample for \underline{h}^g can be drawn in the manner described above, with the channel model order (r^g) being determined using the model order selection routine which will be described later.

A similar analysis for the conditional density $p(\sigma_e^2 | \dots)$ shows that:

$$p(\sigma_e^2|\dots) \propto (\sigma_e^2)^{-\frac{N}{2}} \exp\left[\frac{-E}{2\sigma_e^2}\right] \frac{(\sigma_e^2)^{-(\alpha_e+1)}}{\beta_e \Gamma(\alpha_e)} \exp\left[\frac{-1}{\sigma_e^2 \beta_e}\right] \quad (25)$$

where:

$$E = \mathbf{s}(n)^T \mathbf{s}(n) - 2\mathbf{a}^T \mathbf{S} \mathbf{s}(n) + \mathbf{a}^T \mathbf{S}^T \mathbf{S} \mathbf{a}$$

which can be simplified to give:

$$p(\sigma_e^2|\dots) \propto (\sigma_e^2)^{-\left[\left(\frac{N}{2} + \alpha_e\right) + 1\right]} \exp\left[\frac{-1}{\sigma_e^2} \left(\frac{E}{2} + \frac{1}{\beta_e}\right)\right] \quad (26)$$

which is also an Inverse Gamma distribution having the following parameters:

$$\hat{\alpha}_e = \frac{N}{2} + \alpha_e \quad \text{and} \quad \hat{\beta}_e = \frac{2\beta_e}{2 + \beta_e E} \quad (27)$$

A sample is then drawn from this Inverse Gamma distribution by firstly generating a random number from a uniform distribution and then performing a transformation of random variables using the alpha and beta parameters given in equation (27), to give $(\sigma_e^2)^g$.

A similar analysis for the conditional density $p(\sigma_e^2|\dots)$ reveals that it also is an Inverse Gamma distribution having the following parameters:

$$\hat{\alpha}_e = \frac{N}{2} + \alpha_e \quad \text{and} \quad \hat{\beta}_e = \frac{2\beta_e}{2 + \beta_e E^*} \quad (28)$$

where:

$$E^* = q(n)^T q(n) - 2h^T Y q(n) + h^T Y^T Y h$$

A sample is then drawn from this Inverse Gamma distribution in the manner described above to give $(\sigma_e^2)^g$.

A similar analysis for conditional density $p(\sigma_a^2 | \dots)$ reveals that it too is an Inverse Gamma distribution having the following parameters:

$$\hat{\alpha}_a = \frac{N}{2} + \alpha_a \quad \text{and} \quad \hat{\beta}_a = \frac{2\beta_a}{2 + \beta_a (a - \mu_a)^T (a - \mu_a)} \quad (29)$$

A sample is then drawn from this Inverse Gamma distribution in the manner described above to give $(\sigma_a^2)^g$.

Similarly, the conditional density $p(\sigma_h^2 | \dots)$ is also an Inverse Gamma distribution but having the following parameters:

$$\hat{\alpha}_h = \frac{N}{2} + \alpha_h \quad \text{and} \quad \hat{\beta}_h = \frac{2\beta_h}{2 + \beta_h (h - \mu_h)^T (h - \mu_h)} \quad (30)$$

A sample is then drawn from this Inverse Gamma distribution in the manner described above to give $(\sigma_h^2)^g$.

As those skilled in the art will appreciate, the Gibbs sampler requires an initial transient period to converge to equilibrium (known as burn-in). Eventually, after L iterations, the sample $(\underline{a}^L, k^L, \underline{h}^L, r^L, (\sigma_e^2)^L, (\sigma_\epsilon^2)^L, (\sigma_a^2)^L, (\sigma_h^2)^L, s(n)^L)$ is considered to be a sample from the joint probability density function defined in equation (19). In this embodiment, the Gibbs sampler performs approximately one hundred and fifty (150) iterations on each frame of input speech and discards the samples from the first fifty iterations and uses the rest to give a picture (a set of histograms) of what the joint probability density function defined in equation (19) looks like. From these histograms, the set of AR coefficients (\underline{a}) which best represents the observed speech samples $(y(n))$ from the analogue to digital converter 17 are determined. The histograms are also used to determine appropriate values for the variances and channel model coefficients (\underline{h}) which can be used as the initial values for the Gibbs sampler when it processes the next frame of speech.

Model Order Selection

As mentioned above, during the Gibbs iterations, the model order (k) of the AR filter and the model order (r) of the channel filter are updated using a model order selection routine. In this embodiment, this is performed using a technique derived from "Reversible jump Markov chain Monte Carlo computation", which is described in the paper entitled "Reversible jump Markov chain Monte Carlo Computation and Bayesian model determination" by Peter Green, *Biometrika*, vol 82, pp 711 to 732, 1995.

Figure 4 is a flow chart which illustrates the processing steps performed during this model order selection routine for the AR filter model order (k). As shown, in step s1, a new model order (k_2) is proposed. In this embodiment, the new model order will normally be proposed as $k_2 = k_1 \pm 1$, but occasionally it will be proposed as $k_2 = k_1 \pm 2$ and very occasionally as $k_2 = k_1 \pm 3$ etc. To achieve this, a sample is drawn from a discretised Laplacian density function centred on the current model order (k_1) and with the variance of this Laplacian density function being chosen *a priori* in accordance with the degree of sampling of the model order space that is required.

The processing then proceeds to step s3 where a model

order variable (MO) is set equal to:

$$MO = \max \left\{ \frac{p(\underline{a}_{<1:k_2>}, k_2 | \dots)}{p(\underline{a}_{<1:k_1>}, k_1 | \dots)}, 1 \right\} \quad (31)$$

5 where the ratio term is the ratio of the conditional probability given in equation (21) evaluated for the current AR filter coefficients (\underline{a}) drawn by the Gibbs sampler for the current model order (k_1) and for the proposed new model order (k_2). If $k_2 > k_1$, then the
10 matrix S must first be resized and then a new sample must be drawn from the Gaussian distribution having the mean vector and covariance matrix defined by equations (22) and (23) (determined for the resized matrix S), to provide the AR filter coefficients ($\underline{a}_{<1:k_2>}$) for the new
15 model order (k_2). If $k_2 < k_1$ then all that is required is to delete the last ($k_1 - k_2$) samples from the \underline{a} vector. If the ratio in equation (31) is greater than one, then this implies that the proposed model order (k_2) is better than the current model order whereas if it is less than
20 one then this implies that the current model order is better than the proposed model order. However, since occasionally this will not be the case, rather than deciding whether or not to accept the proposed model order by comparing the model order variable (MO) with a
25 fixed threshold of one, in this embodiment, the model

order variable (MO) is compared, in step s5, with a random number which lies between zero and one. If the model order variable (MO) is greater than this random number, then the processing proceeds to step s7 where the model order is set to the proposed model order (k_2) and a count associated with the value of k_2 is incremented. If, on the other hand, the model order variable (MO) is smaller than the random number, then the processing proceeds to step s9 where the current model order is maintained and a count associated with the value of the current model order (k_1) is incremented. The processing then ends.

This model order selection routine is carried out for both the model order of the AR filter model and for the model order of the channel filter model. This routine may be carried out at each Gibbs iteration. However, this is not essential. Therefore, in this embodiment, this model order updating routine is only carried out every third Gibbs iteration.

Simulation Smoother

As mentioned above, in order to be able to draw samples using the Gibbs sampler, estimates of the raw speech samples are required to generate $\underline{s}(n)$, S and Y which are

used in the Gibbs calculations. These could be obtained from the conditional probability density function $p(\underline{s}(n)|\dots)$. However, this is not done in this embodiment because of the high dimensionality of $\underline{s}(n)$.

5 Therefore, in this embodiment, a different technique is used to provide the necessary estimates of the raw speech samples. In particular, in this embodiment, a "Simulation Smoother" is used to provide these estimates. This Simulation Smoother was proposed by Piet de Jong in
10 the paper entitled "The Simulation Smoother for Time Series Models", Biometrika (1995), vol 82,2, pages 339 to 350. As those skilled in the art will appreciate, the Simulation Smoother is run before the Gibbs Sampler. It is also run again during the Gibbs iterations in order to
15 update the estimates of the raw speech samples. In this embodiment, the Simulation Smoother is run every fourth Gibbs iteration.

In order to run the Simulation Smoother, the model
20 equations defined above in equations (4) and (6) must be written in "state space" format as follows:

$$\begin{aligned}\hat{\underline{s}}(n) &= \tilde{A}.\hat{\underline{s}}(n-1) + \hat{\underline{e}}(n) \\ y(n) &= \underline{h}^T.\hat{\underline{s}}(n-1) + \varepsilon(n)\end{aligned}\tag{32}$$

where

$$\tilde{A} = \begin{bmatrix} a_1 & a_2 & a_3 & \dots & a_k & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \dots & \dots & \dots & 1 & 0 \end{bmatrix}_{rxr}$$

and

$$\hat{s}(n) = \begin{bmatrix} \hat{s}(n) \\ \hat{s}(n-1) \\ \hat{s}(n-2) \\ \vdots \\ \hat{s}(n-r+1) \end{bmatrix}_{rx1} \quad \hat{e}(n) = \begin{bmatrix} \hat{e}(n) \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{rx1}$$

With this state space representation, the dimensionality of the raw speech vectors ($\hat{s}(n)$) and the process noise vectors ($\hat{e}(n)$) do not need to be $N \times 1$ but only have to be as large as the greater of the model orders - k and r . Typically, the channel model order (r) will be larger than the AR filter model order (k). Hence, the vector of raw speech samples ($\hat{s}(n)$) and the vector of process noise ($\hat{e}(n)$) only need to be $rx1$ and hence the dimensionality of the matrix \tilde{A} only needs to be rxr .

The Simulation Smoother involves two stages - a first stage in which a Kalman filter is run on the speech samples in the current frame and then a second stage in

which a "smoothing" filter is run on the speech samples in the current frame using data obtained from the Kalman filter stage. Figure 5 is a flow chart illustrating the processing steps performed by the Simulation Smoother.

As shown, in step s21, the system initialises a time variable t to equal one. During the Kalman filter stage, this time variable is run from $t = 1$ to N in order to process the N speech samples in the current frame being processed in time sequential order. After step s21, the processing then proceeds to step s23, where the following Kalman filter equations are computed for the current speech sample ($y(t)$) being processed:

$$\begin{aligned}
 w(t) &= y(t) - h^T \hat{s}(t) \\
 d(t) &= h^T P(t) h + \sigma_e^2 \\
 k_f(t) &= (\tilde{A} P(t) h) \cdot d(t)^{-1} \\
 \hat{s}(t+1) &= \tilde{A} \hat{s}(t) + k_f(t) \cdot w(t) \\
 L(t) &= \tilde{A} - k_f(t) \cdot h^T \\
 P(t+1) &= \tilde{A} P(t) L(t)^T + \sigma_e^2 \cdot I
 \end{aligned} \tag{33}$$

where the initial vector of raw speech samples ($\hat{s}(1)$) includes raw speech samples obtained from the processing of the previous frame (or if there are no previous frames then $s(i)$ is set equal to zero for $i < 1$); $P(1)$ is the variance of $\hat{s}(1)$ (which can be obtained from the previous

frame or initially can be set to σ_e^2); \underline{h} is the current
 set of channel model coefficients which can be obtained
 from the processing of the previous frame (or if there
 are no previous frames then the elements of \underline{h} can be set
 5 to their expected values - zero); $y(t)$ is the current
 speech sample of the current frame being processed and I
 is the identity matrix. The processing then proceeds to
 step s25 where the scalar values $w(t)$ and $d(t)$ are stored
 together with the rxr matrix $L(t)$ (or alternatively the
 10 Kalman filter gain vector $k_f(t)$ could be stored from
 which $L(t)$ can be generated). The processing then
 proceeds to step s27 where the system determines whether
 or not all the speech samples in the current frame have
 been processed. If they have not, then the processing
 15 proceeds to step s29 where the time variable t is
 incremented by one so that the next sample in the current
 frame will be processed in the same way. Once all N
 samples in the current frame have been processed in this
 way and the corresponding values stored, the first stage
 20 of the Simulation Smoother is complete.

The processing then proceeds to step s31 where the second
 stage of the Simulation Smoother is started in which the
 smoothing filter processes the speech samples in the
 25 current frame in reverse sequential order. As shown, in

step s31 the system runs the following set of smoothing filter equations on the current speech sample being processed together with the stored Kalman filter variables computed for the current speech sample being processed:

$$C(t) = \sigma_e^2 (I - \sigma_e^2 U(t))$$

$$\eta(t) \sim N(0, C(t))$$

$$V(t) = \sigma_e^2 U(t) L(t)$$

$$r(t-1) = h d(t)^{-1} w(t) + L(t)^T r(t) - V(t)^T C(t)^{-1} \eta(t) \quad (34)$$

$$U(t-1) = h d(t)^{-1} h^T + L(t)^T U(t) L(t) + V(t)^T C(t)^{-1} V(t)$$

$$\tilde{e}(t) = \sigma_e^2 r(t) + \eta(t) \quad \text{where } \tilde{e}(t) = [\tilde{e}(t) \ \tilde{e}(t-1) \ \tilde{e}(t-2) \ \dots \ \tilde{e}(t-r+1)]^T$$

$$\hat{s}(t) = \tilde{A} \hat{s}(t-1) + \hat{e}(t) \quad \text{where } \hat{s}(t) = [\hat{s}(t) \ \hat{s}(t-1) \ \hat{s}(t-2) \ \dots \ \hat{s}(t-r+1)]^T$$

$$\text{and } \hat{e}(t) = [\tilde{e}(t) \ 0 \ 0 \ \dots \ 0]^T$$

where $\eta(t)$ is a sample drawn from a Gaussian distribution having zero mean and covariance matrix $C(t)$; the initial vector $r(t=N)$ and the initial matrix $U(t=N)$ are both set to zero; and $\underline{s}(0)$ is obtained from the processing of the previous frame (or if there are no previous frames can be set equal to zero). The processing then proceeds to step s33 where the estimate of the process noise ($\tilde{e}(t)$) for the current speech sample being processed and the estimate of the raw speech sample ($\hat{s}(t)$) for the current speech sample being processed are stored. The processing then proceeds to step s35 where the system determines

whether or not all the speech samples in the current frame have been processed. If they have not, then the processing proceeds to step s37 where the time variable t is decremented by one so that the previous sample in the current frame will be processed in the same way. Once all N samples in the current frame have been processed in this way and the corresponding process noise and raw speech samples have been stored, the second stage of the Simulation Smoother is complete and an estimate of $\underline{s}(n)$ will have been generated.

As shown in equations (4) and (8), the matrix S and the matrix Y require raw speech samples $s(n-N-1)$ to $s(n-N-k+1)$ and $s(n-N-1)$ to $s(n-N-r+1)$ respectively in addition to those in $\underline{s}(n)$. These additional raw speech samples can be obtained either from the processing of the previous frame of speech or if there are no previous frames, they can be set to zero. With these estimates of raw speech samples, the Gibbs sampler can be run to draw samples from the above described probability density functions.

Statistical Analysis Unit - Operation

A description has been given above of the theory underlying the statistical analysis unit 21. A

description will now be given with reference to Figures 6 to 8 of the operation of the statistical analysis unit 21.

5 Figure 6 is a block diagram illustrating the principal components of the statistical analysis unit 21 of this embodiment. As shown, it comprises the above described Gibbs sampler 41, Simulation Smoother 43 (including the
10 Kalman filter 43-1 and smoothing filter 43-2) and model order selector 45. It also comprises a memory 47 which receives the speech samples of the current frame to be processed, a data analysis unit 49 which processes the data generated by the Gibbs sampler 41 and the model order selector 45 and a controller 50 which controls the
15 operation of the statistical analysis unit 21.

As shown in Figure 6, the memory 47 includes a non volatile memory area 47-1 and a working memory area 47-2. The non volatile memory 47-1 is used to store the joint
20 probability density function given in equation (19) above and the equations for the variances and mean values and the equations for the Inverse Gamma parameters given above in equations (22) to (24) and (27) to (30) for the above mentioned conditional probability density functions
25 for use by the Gibbs sampler 41. The non volatile memory

47-1 also stores the Kalman filter equations given above in equation (33) and the smoothing filter equations given above in equation 34 for use by the Simulation Smoother 43.

5

Figure 7 is a schematic diagram illustrating the parameter values that are stored in the working memory area (RAM) 47-2. As shown, the RAM includes a store 51 for storing the speech samples $y_f(1)$ to $y_f(N)$ output by the analogue to digital converter 17 for the current frame (f) being processed. As mentioned above, these speech samples are used in both the Gibbs sampler 41 and the Simulation Smoother 43. The RAM 47-2 also includes a store 53 for storing the initial estimates of the model parameters ($g=0$) and the M samples ($g = 1$ to M) of each parameter drawn from the above described conditional probability density functions by the Gibbs sampler 41 for the current frame being processed. As mentioned above, in this embodiment, M is 100 since the Gibbs sampler 41 performs 150 iterations on each frame of input speech with the first fifty samples being discarded. The RAM 47-2 also includes a store 55 for storing $W(t)$, $d(t)$ and $L(t)$ for $t = 1$ to N which are calculated during the processing of the speech samples in the current frame of speech by the above described Kalman filter 43-1. The

10

15

20

25

RAM 47-2 also includes a store 57 for storing the estimates of the raw speech samples ($\hat{s}_f(t)$) and the estimates of the process noise ($\tilde{e}_f(t)$) generated by the smoothing filter 43-2, as discussed above. The RAM 47-2 also includes a store 59 for storing the model order counts which are generated by the model order selector 45 when the model orders for the AR filter model and the channel model are updated.

Figure 8 is a flow diagram illustrating the control program used by the controller 50, in this embodiment, to control the processing operations of the statistical analysis unit 21. As shown, in step s41, the controller 50 retrieves the next frame of speech samples to be processed from the buffer 19 and stores them in the memory store 51. The processing then proceeds to step s43 where initial estimates for the channel model, raw speech samples and the process noise and measurement noise statistics are set and stored in the store 53. These initial estimates are either set to be the values obtained during the processing of the previous frame of speech or, where there are no previous frames of speech, are set to their expected values (which may be zero). The processing then proceeds to step s45 where the Simulation Smoother 43 is activated so as to provide an

estimate of the raw speech samples in the manner described above. The processing then proceeds to step s47 where one iteration of the Gibbs sampler 41 is run in order to update the channel model, speech model and the process and measurement noise statistics using the raw speech samples obtained in step s45. These updated parameter values are then stored in the memory store 53.

The processing then proceeds to step s49 where the controller 50 determines whether or not to update the model orders of the AR filter model and the channel model. As mentioned above, in this embodiment, these model orders are updated every third Gibbs iteration. If the model orders are to be updated, then the processing proceeds to step s51 where the model order selector 45 is used to update the model orders of the AR filter model and the channel model in the manner described above. If at step s49 the controller 50 determines that the model orders are not to be updated, then the processing skips step s51 and the processing proceeds to step s53. At step s53, the controller 50 determines whether or not to perform another Gibbs iteration. If another iteration is to be performed, then the processing proceeds to decision block s55 where the controller 50 decides whether or not to update the estimates of the raw speech samples ($s(t)$).

If the raw speech samples are not to be updated, then the processing returns to step s47 where the next Gibbs iteration is run.

5 As mentioned above, in this embodiment, the Simulation
Smoother 43 is run every fourth Gibbs iteration in order
to update the raw speech samples. Therefore, if the
controller 50 determines, in step s55 that there has been
four Gibbs iterations since the last time the speech
10 samples were updated, then the processing returns to step
s45 where the Simulation Smoother is run again to provide
new estimates of the raw speech samples ($s(t)$). Once the
controller 50 has determined that the required 150 Gibbs
iterations have been performed, the controller 50 causes
15 the processing to proceed to step s57 where the data
analysis unit 49 analyses the model order counts
generated by the model order selector 45 to determine the
model orders for the AR filter model and the channel
model which best represents the current frame of speech
20 being processed. The processing then proceeds to step
s59 where the data analysis unit 49 analyses the samples
drawn from the conditional densities by the Gibbs sampler
41 to determine the AR filter coefficients (\underline{a}), the
channel model coefficients (\underline{h}), the variances of these
25 coefficients and the process and measurement noise

variances which best represent the current frame of speech being processed. The processing then proceeds to step s61 where the controller 50 determines whether or not there is any further speech to be processed. If there is more speech to be processed, then processing returns to step S41 and the above process is repeated for the next frame of speech. Once all the speech has been processed in this way, the processing ends.

Data Analysis unit

A more detailed description of the data analysis unit 49 will now be given with reference to Figure 9. As mentioned above, the data analysis unit 49 initially determines, in step s57, the model orders for both the AR filter model and the channel model which best represents the current frame of speech being processed. It does this using the counts that have been generated by the model order selector 45 when it was run in step s51. These counts are stored in the store 59 of the RAM 47-2. In this embodiment, in determining the best model orders, the data analysis unit 49 identifies the model order having the highest count. Figure 9a is an exemplary histogram which illustrates the distribution of counts that is generated for the model order (k) of the AR filter model. Therefore, in this example, the data

analysis unit 49 would set the best model order of the AR filter model as five. The data analysis unit 49 performs a similar analysis of the counts generated for the model order (r) of the channel model to determine the best model order for the channel model.

Once the data analysis unit 49 has determined the best model orders (k and r), it then analyses the samples generated by the Gibbs sampler 41 which are stored in the store 53 of the RAM 47-2, in order to determine parameter values that are most representative of those samples.

It does this by determining a histogram for each of the parameters from which it determines the most representative parameter value. To generate the

histogram, the data analysis unit 49 determines the maximum and minimum sample value which was drawn by the Gibbs sampler and then divides the range of parameter values between this minimum and maximum value into a predetermined number of sub-ranges or bins. The data

analysis unit 49 then assigns each of the sample values into the appropriate bins and counts how many samples are allocated to each bin. It then uses these counts to calculate a weighted average of the samples (with the weighting used for each sample depending on the count for the corresponding bin), to determine the most

representative parameter value (known as the minimum mean square estimate (MMSE)). Figure 9b illustrates an example histogram which is generated for the variance (σ_e^2) of the process noise, from which the data analysis unit 49 determines that the variance representative of the sample is 0.3149.

In determining the AR filter coefficients (a_i for $i = 1$ to k), the data analysis unit 49 determines and analyses a histogram of the samples for each coefficient independently. Figure 9c shows an exemplary histogram obtained for the third AR filter coefficient (a_3), from which the data analysis unit 49 determines that the coefficient representative of the samples is -0.4977.

In this embodiment, the data analysis unit 49 only outputs the AR filter coefficients which are passed to the coefficient convertor 23 shown in Figure 2. The remaining parameter values determined by the data analysis unit 49 are stored in the RAM 47-2 for use during the processing of the next frame of speech. As mentioned above, the AR filter coefficients output by the statistical analysis unit 21 are input to the coefficient convertor 23 which converts these coefficients into cepstral coefficients which are then compared with stored

speech models 27 by the speech recognition unit 25 in order to generate a recognition result.

As the skilled reader will appreciate, a speech processing technique has been described above which uses statistical analysis techniques to determine sets of AR filter coefficients representative of an input speech signal. The technique is more robust and accurate than prior art techniques which employ maximum likelihood estimators to determine the AR filter coefficients. This is because the statistical analysis of each frame uses knowledge obtained from the processing of the previous frame. In addition, with the analysis performed above, the model order for the AR filter model is not assumed to be constant and can vary from frame to frame. In this way, the optimum number of AR filter coefficients can be used to represent the speech within each frame. As a result, the AR filter coefficients output by the statistical analysis unit 21 will more accurately represent the corresponding input speech. Further still, since the underlying process model that is used separates the speech source from the channel, the AR filter coefficients that are determined will be more representative of the actual speech and will be less likely to include distortive effects of the channel.

Further still, since variance information is available for each of the parameters, this provides an indication of the confidence of each of the parameter estimates. This is in contrast to maximum likelihood and least square approaches, such as linear prediction analysis, where point estimates of the parameter values are determined.

MULTI SPEAKER MULTI MICROPHONE

A description will now be given of a multi speaker and multi microphone system which uses a similar statistical analysis to separate and model the speech from each speaker. Again, to facilitate understanding, a description will initially be given of a two speaker and two microphone system before generalising to a multi speaker and multi microphone system.

Figure 10 is a schematic block diagram illustrating a speech recognition system which employs a statistical analysis unit embodying the present invention. As shown, the system has two microphones 7-1 and 7-2 which convert, in this embodiment, the speech from two speakers (not shown) into equivalent electrical signals which are passed to a respective filter circuit 15-1 and 15-2. In this embodiment, the filters 15 remove frequencies above

8 kHz since the filtered signals are then converted into corresponding digital signals at a sampling rate of 16 kHz by a respective analogue to digital converter 17-1 and 17-2. The digitized speech samples from the analogue to digital converters 17 are then fed into the buffer 19. The statistical analysis unit 21 analyses the speech within successive frames of the input speech signal from the two microphones. In this embodiment, since there are two microphones there are two sequences of frames which are to be processed. In this embodiment, the two frame sequences are processed together so that the frame of speech from microphone 7-1 at time t is processed with the frame of speech received from the microphone 7-2 at time t . Again, in this embodiment, the frames of speech are non-overlapping and have a duration of 20 ms which, with the 16 kHz sampling rate of the analogue to digital converters 17, results in the statistical analysis unit 21 processing blocks of 640 speech samples (corresponding to two frames of 320 samples).

In order to perform the statistical analysis on the input speech, the analysis unit 21 assumes that there is an underlying process similar to that of the single speaker single microphone system described above. The particular model used in this embodiment is illustrated in Figure

11. As shown, the process is modelled by two speech sources 31-1 and 31-2 which generate, at time $t = n$, raw speech samples $s^1(n)$ and $s^2(n)$ respectively. Again, in this embodiment, each of the speech sources 31 is modelled by an auto aggressive (AR) process. In other words, there will be a respective equation (1) for each of the sources 31-1 and 31-2, thereby defining two unknown AR filter coefficient vectors \underline{a}^1 and \underline{a}^2 , each having a respective model order k^1 and k^2 . These source models will also have a respective process noise component $e^1(n)$ and $e^2(n)$.

As shown in Figure 11, the model also assumes that the speech generated by each of the sources 31 is received by both microphones 7. There is therefore a respective channel 33-11 to 33-22 between each source 31 and each microphone 7. There is also a respective measurement noise component $\varepsilon^1(n)$ and $\varepsilon^2(n)$ added to the signal received by each microphone. Again, in this embodiment, the statistical analysis unit 21 models each of the channels by a moving average (MA) filter. Therefore, the signal received from microphone 7-1 at time $t = n$ is given by:

$$\begin{aligned}
 y^1(n) = & h_{110}s^1(n) + h_{111}s^1(n-1) + h_{112}s^1(n-2) + \dots + h_{11r_1}s^1(n-r_{11}) \\
 & + h_{210}s^2(n) + h_{211}s^2(n-1) + h_{212}s^2(n-2) + \dots + h_{21r_{21}}s^2(n-r_{21}) + \varepsilon^1(n)
 \end{aligned}
 \tag{35}$$

where, for example, h_{112} is the channel filter coefficient of the channel between the first source 31-1 and the microphone 7-1 at time $t = 2$; and r_{21} is the model order of the channel between the second speech source 31-2 and the microphone 7-1. A similar equation will exist to represent the signal received from the other microphone 7-2.

In this embodiment, the statistical analysis unit 21 aims to determine values for the AR filter coefficients for the two speech sources, which best represent the observed signal samples from the two microphones in the current frame being processed. It does this, by determining the AR filter coefficients for the two speakers (\underline{a}^1 and \underline{a}^2) that maximise the joint probability density function of the speech models, channel models, raw speech samples and the noise statistics given the observed signal samples output from the two analogue to digital converters 17-1 and 17-2, i.e. by determining:

$$\max_{\underline{a}^1, \underline{a}^2} \left\{ p(\underline{a}^1, \underline{a}^2, k^1, k^2, h_{11}, h_{12}, h_{21}, h_{22}, r_{11}, r_{12}, r_{21}, r_{22}, \sigma_{e_1}^2, \sigma_{e_2}^2, \sigma_{\epsilon_1}^2, \sigma_{\epsilon_2}^2, s^1(n), s^2(n) | y^1(n), y^2(n)) \right\} \quad (36)$$

As those skilled in the art will appreciate, this is almost an identical problem to the single speaker single microphone system described above, although with more

parameters. Again, to calculate this, the above probability is rearranged using Bayes law to give an equation similar to that given in equation (10) above. The only difference is that there will be many more joint probability density functions on the numerator. In particular, the joint probability density functions which will need to be considered in this embodiment are:

$$\begin{aligned}
 & p(\mathbf{y}^1(n) | \mathbf{s}^1(n), \mathbf{s}^2(n), \mathbf{h}_{11}, \mathbf{h}_{21}, \mathbf{r}_{11}, \mathbf{r}_{21}, \sigma_{e1}^2) \\
 & p(\mathbf{y}^2(n) | \mathbf{s}^1(n), \mathbf{s}^2(n), \mathbf{h}_{12}, \mathbf{h}_{22}, \mathbf{r}_{12}, \mathbf{r}_{22}, \sigma_{e2}^2) \\
 & p(\mathbf{s}^1(n) | \mathbf{a}^1, \mathbf{k}^1, \sigma_{e1}^2) \quad p(\mathbf{s}^2(n) | \mathbf{a}^2, \mathbf{k}^2, \sigma_{e2}^2) \\
 & p(\mathbf{a}^1 | \mathbf{k}^1, \sigma_{a1}^2, \mu_{a1}) \quad p(\mathbf{a}^2 | \mathbf{k}^2, \sigma_{a2}^2, \mu_{a2}) \\
 & p(\mathbf{h}_{11} | \mathbf{r}_{11}, \sigma_{h11}^2, \mu_{h11}) \quad p(\mathbf{h}_{12} | \mathbf{r}_{12}, \sigma_{h12}^2, \mu_{h12}) \\
 & p(\mathbf{h}_{21} | \mathbf{r}_{21}, \sigma_{h21}^2, \mu_{h21}) \quad p(\mathbf{h}_{22} | \mathbf{r}_{22}, \sigma_{h22}^2, \mu_{h22}) \\
 & P(\sigma_{a1}^2 | \alpha_{a1}, \beta_{a1}) \quad P(\sigma_{a2}^2 | \alpha_{a2}, \beta_{a2}) \quad p(\sigma_{e1}^2) \quad p(\sigma_{e2}^2) \\
 & P(\sigma_{h11}^2 | \alpha_{h11}, \beta_{h11}) \quad P(\sigma_{h12}^2 | \alpha_{h12}, \beta_{h12}) \quad P(\sigma_{h21}^2 | \alpha_{h21}, \beta_{h21}) \\
 & P(\sigma_{h22}^2 | \alpha_{h22}, \beta_{h22}) \quad p(\mathbf{k}^1) \quad p(\mathbf{k}^2) \quad p(\mathbf{r}_{11}) \quad p(\mathbf{r}_{12}) \quad p(\mathbf{r}_{21}) \quad p(\mathbf{r}_{22})
 \end{aligned}$$

Since the speech sources and the channels are independent of each other, most of these components will be the same as the probability density functions given above for the single speaker single microphone system. This is not the case, however, for the joint probability density functions for the vectors of speech samples ($\mathbf{y}^1(n)$ and $\mathbf{y}^2(n)$) out from the analogue to digital converters 17,

since these signals include components from both the speech sources. The joint probability density function for the speech samples output from analogue to digital converter 17-1 will now be described in more detail.

5

$$p(\mathbf{y}^1(n) | \mathbf{s}^1(n), \mathbf{s}^2(n), \mathbf{h}_{11}, \mathbf{h}_{21}, \mathbf{r}_{11}, \mathbf{r}_{21}, \sigma_{\epsilon^1}^2)$$

Considering all the speech samples output from the analogue to digital converter 17-1 in a current frame being processed (and with h_{110} and h_{210} being set equal to one), gives:

10

$$\mathbf{\epsilon}^1(n) = \mathbf{q}^1(n) - \begin{bmatrix} Y_1 & Y_2 \end{bmatrix} \cdot \begin{bmatrix} h_{11} \\ \vdots \\ h_{21} \end{bmatrix} \quad (37)$$

where

15

$$\mathbf{h}_{11} = \begin{bmatrix} h_{111} \\ h_{112} \\ h_{113} \\ \vdots \\ h_{11r_{11}} \end{bmatrix}_{r_{11} \times 1} \quad \mathbf{h}_{21} = \begin{bmatrix} h_{211} \\ h_{212} \\ h_{213} \\ \vdots \\ h_{21r_{21}} \end{bmatrix}_{r_{21} \times 1} \quad \mathbf{q}^1(n) = \begin{bmatrix} q^1(n) \\ q^1(n-1) \\ q^1(n-2) \\ \vdots \\ q^1(n-N+1) \end{bmatrix}_{N \times 1} \quad \mathbf{\epsilon}^1(n) = \begin{bmatrix} \epsilon^1(n) \\ \epsilon^1(n-1) \\ \epsilon^1(n-2) \\ \vdots \\ \epsilon^1(n-N+1) \end{bmatrix}_{N \times 1}$$

20

and

$$\mathbf{Y}_1 = \begin{bmatrix} s^1(n-1) & s^1(n-2) & \dots & s^1(n-r_{11}) \\ s^1(n-2) & s^1(n-3) & \dots & s^1(n-r_{11}-1) \\ s^1(n-3) & s^1(n-4) & \dots & s^1(n-r_{11}-2) \\ \vdots & \vdots & \ddots & \vdots \\ s^1(n-N) & s^1(n-N-1) & \dots & s^1(n-r_{11}-N+1) \end{bmatrix}_{N \times r_{11}}$$

25

$$Y_2 = \begin{bmatrix} s^2(n-1) & s^2(n-2) & \dots & s^2(n-r_{21}) \\ s^2(n-2) & s^2(n-3) & \dots & s^2(n-r_{21}-1) \\ s^2(n-3) & s^2(n-4) & \dots & s^2(n-r_{21}-2) \\ \vdots & \vdots & \ddots & \vdots \\ s^2(n-N) & s^2(n-N-1) & \dots & s^2(n-r_{21}-N+1) \end{bmatrix}_{N \times r_{21}}$$

and $q^1(n) = y^1(n) - s^1(n) - s^2(n)$.

As in the single speaker single microphone system described above, the joint probability density function for the speech samples ($y^1(n)$) output from the analogue to digital converter 17-1 is determined from the joint probability density function for the associated measurement noise ($\sigma_{e_1}^2$) using equation (14) above. Again, the Jacobean will be one and the resulting joint probability density function will have the following form:

$$p(y^1(n) | s^1(n), s^2(n), h_{11}, h_{21}, r_{11}, r_{21}, \sigma_{e_1}^2) = (2\pi\sigma_{e_1}^2)^{-\frac{N}{2}} \exp \left[\frac{-1}{2\sigma_{e_1}^2} \left(q^1(n)^T q^1(n) - 2q^1(n) \begin{bmatrix} Y_1 & Y_2 \end{bmatrix} \cdot \begin{bmatrix} h_{11} \\ \vdots \\ h_{21} \end{bmatrix} + \begin{bmatrix} h_{11}^T & h_{21}^T \end{bmatrix} \cdot \begin{bmatrix} Y_1^T Y_1 & Y_2^T Y_1 \\ Y_1^T Y_2 & Y_2^T Y_2 \end{bmatrix} \begin{bmatrix} h_{11} \\ \vdots \\ h_{21} \end{bmatrix} \right) \right] \quad (38)$$

As those skilled in the art will appreciate, this is a

Gaussian distribution as before. In this embodiment, the statistical analysis unit 21 assumes that the raw speech data which passes through the two channels to the microphone 7-1 are independent of each other. This allows the above Gaussian distribution to be simplified since the cross components $Y_1^T Y_2$ and $Y_2^T Y_1$ can be assumed to be zero. This gives:

$$\begin{aligned}
 & p(y^1(n) | s^1(n), s^2(n), h_{11}, h_{21}, r_{11}, r_{21}, \sigma_{e_1}^2) \\
 & \propto (2\pi\sigma_{e_1}^2)^{-\frac{N}{2}} \exp \left[\frac{-1}{2\sigma_{e_1}^2} \left(-2h_{11}^T Y_1^T q^1(n) + h_{11}^T Y_1^T Y_1 h_{11} \right) \right] \\
 & \quad + \exp \left[\frac{-1}{2\sigma_{e_1}^2} \left(-2h_{21}^T Y_2^T q^1(n) + h_{21}^T Y_2^T Y_2 h_{21} \right) \right]
 \end{aligned} \tag{39}$$

which is a product of two Gaussians, one for each of the two channels to the microphone 7-1. Note also that the initial term $q^1(n)^T q^1(n)$ has been ignored, since this is just a constant and will therefore only result in a corresponding scaling factor to the probability density function. This simplification is performed in this embodiment, since it is easier to draw a sample from each of the two Gaussians given in equation (39) individually rather than having to draw a single sample of both channels jointly from the larger Gaussian defined by equation (38).

The Gibbs sampler is then used to draw samples from the

5
10

15

20

$$\tilde{A}^{<1:2>} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & . & . & a_1 k^1 & : & & & & & & & & \\ 1 & 0 & 0 & . & . & 0 & : & & & & & & & & \\ 0 & 1 & 0 & . & . & 0 & : & & & 0 & & & & & \\ . & & & & & & : & & & & & & & & \\ . & & & & & 0 & : & & & & & & & & \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \\ & & & & & & : & a_{21} & a_{22} & a_{23} & . & . & a_2 k^2 & & \\ & & & & & & : & 1 & 0 & 0 & . & . & 0 & & \\ & & 0 & & & & : & 0 & 1 & 0 & . & . & 0 & & \\ & & & & & & : & . & & & . & & & & \\ & & & & & & : & . & & & & & 0 & & \end{bmatrix}_{mxm}$$

5

$$\hat{\mathbf{s}}^{<1:2>}(n) = \begin{bmatrix} \hat{s}^1(n) \\ \hat{s}^1(n-1) \\ \hat{s}^1(n-2) \\ \vdots \\ \hat{s}^1(n-r_{11}+1) \\ \dots \\ \hat{s}^2(n) \\ \hat{s}^2(n-1) \\ \hat{s}^2(n-2) \\ \vdots \\ \hat{s}^2(n-r_{21}+1) \end{bmatrix}_{mx1} \quad \hat{\mathbf{e}}^{<1:2>}(n) = \begin{bmatrix} \hat{e}_1(n) \\ 0 \\ 0 \\ \vdots \\ 0 \\ \dots \\ \hat{e}_2(n) \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{mx1} \quad B = \begin{bmatrix} \sigma_{e_1}^2 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ \dots & \dots \\ 0 & \sigma_{e_2}^2 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}_{mx2}$$

10

$$H^{<1:2>} = \begin{bmatrix} h_{111} & h_{112} & h_{113} & \dots & h_{11r_{11}} & \vdots & h_{211} & h_{212} & h_{213} & \dots & h_{21r_{21}} \\ \dots & \dots & \dots & \dots & \dots & \vdots & \dots & \dots & \dots & \dots & \dots \\ h_{121} & h_{122} & h_{123} & \dots & h_{12r_{12}} & \vdots & h_{221} & h_{222} & h_{223} & \dots & h_{22r_{22}} \end{bmatrix}_{2xm}^T$$

15

and

$$\mathbf{y}^{<1:2>}(n) = \begin{bmatrix} y^1(n) \\ \dots \\ y^2(n) \end{bmatrix} \quad D = \begin{bmatrix} \sigma_{e_1}^2 & 0 \\ 0 & \sigma_{e_2}^2 \end{bmatrix} \quad \mathbf{e}^{<1:2>}(n) = \begin{bmatrix} \epsilon^1(n) \\ \dots \\ \epsilon^2(n) \end{bmatrix}$$

20

where m is the larger of the AR filter model orders and the MA filter model orders. Again, this results in slightly more complicated Kalman filter equations and smoothing filter equations and these are given below for completeness.

25

Kalman filter equations

$$\mathbf{w}(t) = \mathbf{y}^{<1:2>}(t) - \mathbf{H}^{<1:2>T} \hat{\mathbf{x}}^{<1:2>}(t)$$

$$\mathbf{D}(t) = \mathbf{H}^{<1:2>T} \mathbf{P}(t) \mathbf{H}^{<1:2>} + \mathbf{D}_e \cdot \mathbf{D}_e^T$$

$$\mathbf{K}_f(t) = (\tilde{\mathbf{A}}^{<1:2>} \mathbf{P}(t) \mathbf{H}^{<1:2>}) \cdot \mathbf{D}(t)^{-1}$$

$$\hat{\mathbf{x}}^{<1:2>}(t+1) = \tilde{\mathbf{A}}^{<1:2>} \hat{\mathbf{x}}^{<1:2>}(t) + \mathbf{K}_f(t) \cdot \mathbf{w}(t) \quad (41)$$

$$\mathbf{L}(t) = \tilde{\mathbf{A}}^{<1:2>} - \mathbf{K}_f(t) \cdot \mathbf{H}^{<1:2>T}$$

$$\mathbf{P}(t+1) = \tilde{\mathbf{A}}^{<1:2>} \mathbf{P}(t) \mathbf{L}(t)^T + \mathbf{D}_e \cdot \mathbf{D}_e^T$$

Smoothing Filter Equations

$$\mathbf{C}(t) = \mathbf{B} \cdot \mathbf{B}^T - \mathbf{B} \cdot \mathbf{B}^T \cdot \mathbf{U}(t) \mathbf{B} \cdot \mathbf{B}^T$$

$$\mathbf{\eta}(t) \sim N(0, \mathbf{C}(t))$$

$$\mathbf{V}(t) = \mathbf{B} \mathbf{B}^T \mathbf{U}(t) \mathbf{L}(t)$$

$$\mathbf{r}(t-1) = \mathbf{H}^{<1:2>} \mathbf{D}(t)^{-1} \mathbf{w}(t) + \mathbf{L}(t)^T \mathbf{r}(t) - \mathbf{V}(t)^T \mathbf{C}(t)^{-1} \mathbf{\eta}(t)$$

$$\mathbf{U}(t-1) = \mathbf{H}^{<1:2>} \mathbf{D}(t)^{-1} \mathbf{H}^{<1:2>T} + \mathbf{L}(t)^T \mathbf{U}(t) \mathbf{L}(t) + \mathbf{V}(t)^T \mathbf{C}(t)^{-1} \mathbf{V}(t) \quad (42)$$

$$\tilde{\mathbf{e}}^{<1:2>}(t) = \mathbf{B} \cdot \mathbf{B}^T \mathbf{r}(t) + \mathbf{\eta}(t)$$

$$\text{where } \tilde{\mathbf{e}}^{<1:2>}(t) = [\tilde{e}_1(t) \ \tilde{e}_1(t-1) \ \dots \ \tilde{e}_1(t-r+1) : \tilde{e}_2(t) \ \tilde{e}_2(t-1) \ \dots \ \tilde{e}_2(t-r+1)]^T$$

$$\hat{\mathbf{x}}^{<1:2>}(t) = \tilde{\mathbf{A}}^{<1:2>} \hat{\mathbf{x}}^{<1:2>}(t-1) + \tilde{\mathbf{e}}^{<1:2>}(t)$$

$$\text{where } \tilde{\mathbf{e}}^{<1:2>}(t) = [\tilde{e}^{<1:2>}(t) \ 0 \ \dots \ 0 : \tilde{\mathbf{e}}^{<1:2>}(t) = [\tilde{e}^{<1:2>}(t) \ 0 \ \dots \ 0]^T$$

The processing steps performed by the statistical analysis unit 21 for this two speaker two microphone system are the same as those used in the single speaker single microphone system described above with reference to Figures 8 and 9 and will not, therefore, be described again.

In the above two speaker two microphone system, the system assumed that there were two speakers. In a general system, the number of speakers at any given time will be unknown. Figure 12 is a block diagram illustrating a multi-speaker multi-microphone speech recognition system. As shown in Figure 12, the system comprises a plurality of microphones 7-1 to 7-j, each of which receives speech signals from an unknown number of speech sources (not shown). The corresponding electrical signals output by the microphones 7 are then passed through a respective filter 15 and then digitized by a respective analogue to digital converter 17. The digitized speech signals from each of the microphones 7 are then stored in the buffer 19 as before. As shown in Figure 12, the speech stored within the buffer 19 is fed into a plurality (m) of statistical analysis units 21. Each of the statistical analysis units is programmed to apply the current frame of speech samples to the

following probability density function and to then draw samples from it in the manner described above:

5

$$\begin{aligned}
 & \prod_{j=1}^{N_{SEN}} \left[(2\pi\sigma_{e_j}^2)^{-\frac{N}{2}} \exp \left[\frac{-1}{2\sigma_{e_j}^2} \left(-2\mathbf{h}_{<1:Z,J>}^T \mathbf{Y}_{<1:Z>}^T \mathbf{y}^j(n) + \mathbf{h}_{<1:Z,J>}^T \mathbf{Y}_{<1:Z>}^T \mathbf{Y}_{<1:Z>} \mathbf{h}_{<1:Z,J>} \right) \right] \right] \\
 & \times \prod_{i=1}^Z \left[(2\pi\sigma_{e_i}^2)^{-\frac{N}{2}} \exp \left[\frac{-1}{2\sigma_{e_i}^2} \left(\mathbf{z}^i(n)^T \mathbf{z}^i(n) - 2\mathbf{a}^{iT} \mathbf{S}^i \mathbf{z}^i(n) + \mathbf{a}^{iT} \mathbf{S}^{iT} \mathbf{S}^i \mathbf{a}^i \right) \right] \right] \\
 & \times \prod_{i=1}^Z \left[(2\pi\sigma_{a_i}^2)^{-\frac{N}{2}} \exp \left[\frac{-(\mathbf{a}^i - \boldsymbol{\mu}_{a_i})^T (\mathbf{a}^i - \boldsymbol{\mu}_{a_i})}{2\sigma_{a_i}^2} \right] \right] \\
 & \times \prod_{j=1}^{N_{SEN}} \left[\prod_{i=1}^Z \left[(2\pi\sigma_{h_{ij}}^2)^{-\frac{N}{2}} \exp \left[\frac{-(\mathbf{h}_{ij} - \boldsymbol{\mu}_{h_{ij}})^T (\mathbf{h}_{ij} - \boldsymbol{\mu}_{h_{ij}})}{2\sigma_{h_{ij}}^2} \right] \right] \right] \\
 & \times \prod_{i=1}^Z \left[\frac{(\sigma_{a_i}^2)^{-(\alpha_{a_i}+1)}}{\beta_{a_i} \Gamma(\alpha_{a_i})} \exp \left[\frac{-1}{\sigma_{a_i}^2 \beta_{a_i}} \right] \right] \\
 & \times \prod_{j=1}^{N_{SEN}} \left[\prod_{i=1}^Z \left[\frac{(\sigma_{h_{ij}}^2)^{-(\alpha_{h_{ij}}+1)}}{\beta_{h_{ij}} \Gamma(\alpha_{h_{ij}})} \exp \left[\frac{-1}{\sigma_{h_{ij}}^2 \beta_{h_{ij}}} \right] \right] \right] \\
 & \times \prod_{i=1}^Z \left[\frac{(\sigma_{e_i}^2)^{-(\alpha_{e_i}+1)}}{\beta_{e_i} \Gamma(\alpha_{e_i})} \exp \left[\frac{-1}{\sigma_{e_i}^2 \beta_{e_i}} \right] \right] \\
 & \times \prod_{j=1}^{N_{SEN}} \left[\frac{(\sigma_{e_j}^2)^{-(\alpha_{e_j}+1)}}{\beta_{e_j} \Gamma(\alpha_{e_j})} \exp \left[\frac{-1}{\sigma_{e_j}^2 \beta_{e_j}} \right] \right]
 \end{aligned}$$

25

(43)

where N_{SEN} is the number of microphones 7 and Z is the number of speakers (which is different for each of the analysis units 21 and is set by a model comparison unit 64). In this way, each of the analysis units 21 performs a similar analysis using the same input data (the speech samples from the microphones) but assumes that the input data was generated by a different number of speakers. For example, statistical analysis unit 21-1 may be programmed to assume that there are three speakers currently speaking whereas statistical analysis unit 21-2 may be programmed to assume that there are five speakers currently speaking etc.

During the processing of each frame of speech by the statistical analysis units 21, some of the parameter samples drawn by the Gibbs sampler are supplied to the model comparison unit 64 so that it can identify the analysis unit that models best the speech in the current frame being processed. In this embodiment samples from every fifth Gibbs iteration are output to the model comparison unit 64 for this determination to be made. After each of the analysis units has finished sampling the above probability density function, it determines the mean AR filter coefficients for the programmed number of speakers in the manner described above and outputs these

to a selector unit 62. At the same time, after the model comparison unit 64 has determined the best analysis unit, it passes a control signal to the selector unit 62 which causes the AR filter coefficients output by this analysis unit 21 to be passed to the speech recognition unit 25 for comparison with the speech models 27. In this embodiment, the model comparison unit 64 is also arranged to reprogram each of the statistical analysis units 21 after the processing of each frame has been completed, so that the number of speakers that each of the analysis units is programmed to model is continuously adapted. In this way, the system can be used in, for example, a meeting where the number of participants speaking at any one time may vary considerably.

Figure 13 is a flow diagram illustrating the processing steps performed in this embodiment, by each of the statistical analysis units 21. As can be seen from a comparison of Figure 13 with Figure 8, the processing steps employed are substantially the same as in the above embodiment, except for the additional steps S52, S54 and S56. A description of these steps will now be given. As shown in Figure 13, if step s54 determines that another Gibbs iteration is to be run, then the processing proceeds to step S52 where each of the

statistical analysis units 21-1 determines whether or not to send the parameter samples from the last Gibbs iteration to the model comparison unit 64. As mentioned above, the model comparison unit 64 compares the samples generated by the analysis units every fifth Gibbs iteration. Therefore, if the samples are to be compared, then the processing proceeds to step S54 where each of the statistical analysis units 21-1 sends the current set of parameter samples to the model comparison unit 64. The processing then proceeds to step S55 as before. Once the analysis units 21 have completed the sampling operation for the current frame, the processing then proceeds to step S56 where each of the statistical analysis units 21-1 informs the model comparison unit 64 that it has completed the Gibbs iterations for the current frame before proceeding to step S57 as before.

The processing steps performed by the model comparison unit 64 in this embodiment will now be described with reference to Figures 14 and 15. As shown, Figure 14 is a flow chart and illustrates the processing steps performed by the model comparison unit 64 when it receives the samples from each of the statistical analysis units 21 during the Gibbs iterations. As shown, in step S71, the model comparison unit 64 uses the

samples received from each of the statistical analysis units 21 to evaluate the probability density function given in equation (43). The processing then proceeds to step S73 where the model comparison unit 64 compares the evaluated probability density functions to determine which statistical analysis unit gives the highest evaluation. The processing then proceeds to step S75 where the model comparison unit 64 increments a count associated with the statistical analysis unit 21 having the highest evaluation. The processing then ends.

Once all the statistical analysis units 21 have carried out all the Gibbs iterations for the current frame of speech being processed, the model comparison unit performs the processing steps shown in Figure 15. In particular, at step S81, the model comparison unit 64 analyses the accumulated counts associated with each of the statistical analysis units, to determine the analysis unit having the highest count. The processing then proceeds to step S83 where the model comparison unit 64 outputs a control signal to the selector unit 62 in order to cause the AR filter coefficients generated by the statistical analysis unit having the highest count to be passed through the selector 62 to the speech recognition unit 25. The processing then proceeds to step S85 where

the model comparison unit 64 determines whether or not it needs to adjust the settings of each of the statistical analysis units 21, and in particular to adjust the number of speakers that each of the statistical analysis units assumes to be present within the speech.

As those skilled in the art will appreciate, a multi speaker multi microphone speech recognition has been described above. This system has all the advantages described above for the single speaker single microphone system. It also has the further advantages that it can simultaneously separate and model the speech from a number of sources. Further, there is no limitation on the physical separation of the sources relative to each other or relative to the microphones. Additionally, the system does not need to know the physical separation between the microphones and it is possible to separate the signals from each source even where the number of microphones is fewer than the number of sources.

Alternative Embodiments

In the above embodiment, the statistical analysis unit was used as a pre-processor for a speech recognition system in order to generate AR coefficients representative of the input speech. It also generated a

number of other parameter values (such as the process noise variances and the channel model coefficients), but these were not output by the statistical analysis unit. As those skilled in the art will appreciate, the AR coefficients and some of the other parameters which are calculated by the statistical analysis unit can be used for other purposes. For example, Figure 16 illustrates a speech recognition system which is similar to the speech recognition system shown in Figure 10 except that there is no coefficient converter since the speech recognition unit 25 and speech models 27 are AR coefficient based. The speech recognition system shown in Figure 16 also has an additional speech detection unit 61 which receives the AR filter coefficients (a) together with the AR filter model order (k) generated by the statistical analysis unit 21 and which is operable to determine from them when speech is present within the signals received from the microphones 7. It can do this, since the AR filter model orders and the AR filter coefficient values will be larger during speech than when there is no speech present. Therefore, by comparing the AR filter model order (k) and/or the AR filter coefficient values with appropriate threshold values, the speech detection unit 61 can determine whether or not speech is present within the input signal. When the

speech detection unit 61 detects the presence of speech, it outputs an appropriate control signal to the speech recognition unit 25 which causes it to start processing the AR coefficients it receives from the statistical analysis unit 21. Similarly, when the speech detection unit 61 detects the end of speech, it outputs an appropriate control signal to the speech recognition unit 25 which causes it to stop processing the AR coefficients it receives from the statistical analysis unit 21.

In the above embodiments, a speech recognition system was described having a particular speech pre-processing front end which performed a statistical analysis of the input speech. As the those skilled in the art will appreciate, this pre-processing can be used in speech processing systems other than speech recognition systems. For example, as shown in Figure 17, the statistical analysis unit 21 may form a front end to a speaker verification system 65. In this embodiment, the speaker verification system 65 compares the sequences of AR filter coefficients for the different speakers output by the statistical analysis unit 21 with pre-stored speaker models 67 to determine whether or not the received speech corresponds to known users.

Figure 18 illustrates another application for the statistical analysis unit 21. In particular, Figure 18 shows an acoustic classification system. The statistical analysis unit 21 is used to generate the AR filter coefficients for each of a number of acoustic sources (which may or may not be speech) in the manner described above. The coefficients are then passed to an acoustic classification system 66 which compares the AR coefficients of each source with pre-stored acoustic models 68 to generate a classification result. Such a system may be used, for example, to distinguish and identify, for example, percussion sounds, woodwind sounds, brass sounds as well as speech.

Figure 19 illustrates another application for the statistical analysis unit 21. In particular, Figure 19 shows a speech encoding and transmission system. The statistical analysis unit 21 is used to generate the AR filter coefficients for each speaker in the manner described above. These coefficients are then passed to a channel encoder which encodes the sequences of AR filter coefficients so that they are in a more suitable form for transmission through a communications channel. The encoded AR filter coefficients are then passed to a transmitter 73 where the encoded data is used to modulate

a carrier signal which is then transmitted to a remote receiver 75. The receiver 75 demodulates the received signal to recover the encoded data which is then decoded by a decoder 76. The sequences of AR filter coefficients output by the decoder are then either passed to a speech recognition unit 77 which compares the sequences of AR filter coefficients with stored reference models (not shown) to generate a recognition result or to a speech synthesis unit 79 which re-generates the speech and outputs it via a loudspeaker 81. As shown, prior to application to the speech synthesis unit 79, the sequences of AR filter coefficients may also pass through an optional processing unit 83 (shown in phantom) which can be used to manipulate the characteristics of the speech that is synthesised. One of the significant advantages of using the statistical analysis unit described above is that the model orders for the AR filter models are not assumed to be constant and will vary from frame to frame. In this way, the optimum number of AR filter coefficients will be used to represent the speech from each speaker within each frame. In contrast, with linear prediction analysis, the number of AR filter coefficients is assumed to be constant and hence the prior art techniques tend to over parameterise the speech in order to ensure that information is not

lost. As a result, with the statistical analysis described above, the amount of data which has to be transmitted from the transmitter to the receiver will be less than with the prior art systems which assume a fixed size of AR filter model.

Figure 20 shows another system which uses the statistical analysis unit 21 described above. The system shown in Figure 20 automatically generates voice annotation data for adding to a data file. The system may be used, for example, to generate voice annotation data for a meeting involving a number of participants, with the data file 91 being a recorded audio file of the meeting. In use, as the meeting progresses, the speech signals received from the microphones is processed by the statistical analysis unit 21 to separate the speech signals from each of the participants. Each participant's speech is then tagged with an identifier identifying who is speaking and then passed to a speech recognition unit 97, which generates words and/or phoneme data for each speaker. This word and/or phoneme data is then passed to a data file annotation unit 99, which annotates the data file 91 with the word and/or phoneme data and then stores the annotated data file in a database 101. In this way, subsequent to the meeting, a user can search the data

file 91 for a particular topic that was discussed at the meeting by a particular participant.

In addition, in this embodiment, the statistical analysis unit 21 also outputs the variance of the AR filter coefficients for each of the speakers. This variance information is passed to a speech quality assessor 93 which determines from this variance data, a measure of the quality of each participant's speech. As those skilled in the art will appreciate, in general, when the input speech is of a high quality (i.e. not disturbed by high levels of background noise), this variance should be small and where there are high levels of noise, this variance should be large. The speech quality assessor 93 then outputs this quality indicator to the data file annotation unit 99 which annotates the data file 91 with this speech quality information.

As the those skilled in the art will appreciate, these speech quality indicators which are stored with the data file are useful for subsequent retrieval operations. In particular, when the user wishes to retrieve a data file 91 from the database 101 (using a voice query), it is useful to know the quality of the speech that was used to annotate the data file and/or the quality of the voice

retrieval query used to retrieve the data file, since this will affect the retrieval performance. In particular if the voice annotation is of a high quality and the user's retrieval query is also of a high quality, then a stringent search of the database 101 can be performed, in order to reduce the amount of false identifications. In contrast, if the original voice annotation is of a low quality or if the user's retrieval query is of a low quality, then a less stringent search of the database 101 can be performed to give a higher chance of retrieving the correct data file 91.

In addition to using the variance of the AR filter coefficients as an indication of the speech quality, the variance (σ_e^2) of the process noise is also a good measure of the quality of the input speech, since this variance is also measure of the energy in the process noise. Therefore, the variance of the process noise can be used in addition to or instead of the variance of the AR filter coefficients to provide the measure of quality of the input speech.

In the embodiment described above with reference to Figure 16, the statistical analysis unit 21 may be used solely for providing information to the speech detection

unit 61 and a separate speech preprocessor may be used to parameterise the input speech for use by the speech recognition unit 25. However, such separate parameterisation of the input speech is not preferred because of the additional processing overhead involved.

The above embodiments have described a statistical analysis technique for processing signals received from a number of microphones in response to speech signals generated by a plurality of speakers. As those skilled in the art will appreciate, the statistical analysis technique described above may be employed in fields other than speech and/or audio processing. For example, the system may be used in fields such as data communications, sonar systems, radar systems etc.

In the first embodiment described above, the AR filter coefficients output by the statistical analysis unit 21 were converted into cepstral coefficients since the speech recognition unit used in the first embodiment was a cepstral based system. As those skilled in the art will appreciate, if the speech recognition system is designed to work with other spectral coefficients, then the coefficient converter 23 may be arranged to convert the AR filter coefficients into the appropriate spectral

parameters. Alternatively still, if the speech recognition system is designed to operate with AR coefficients, then the coefficient converter 23 is unnecessary.

5

In the above embodiments, Gaussian and Inverse Gamma distributions were used to model the various prior probability density functions of equation (19). As those skilled in the art of statistical analysis will appreciate, the reason these distributions were chosen is that they are conjugate to one another. This means that each of the conditional probability density functions which are used in the Gibbs sampler will also either be Gaussian or Inverse Gamma. This therefore simplifies the task of drawing samples from the conditional probability densities. However, this is not essential. The noise probability density functions could be modelled by Laplacian or student-t distributions rather than Gaussian distributions. Similarly, the probability density functions for the variances may be modelled by a distribution other than the Inverse Gamma distribution. For example, they can be modelled by a Rayleigh distribution or some other distribution which is always positive. However, the use of probability density functions that are not conjugate will result in increased

10

15

20

25

complexity in drawing samples from the conditional densities by the Gibbs sampler.

5 Additionally, whilst the Gibbs sampler was used to draw
samples from the probability density function given in
equation (19), other sampling algorithms could be used.
For example the Metropolis-Hastings algorithm (which is
reviewed together with other techniques in a paper
entitled "Probabilistic inference using Markov chain
10 Monte Carlo methods" by R. Neal, Technical Report CRG-
TR-93-1, Department of Computer Science, University of
Toronto, 1993) may be used to sample this probability
density.

15 In the above embodiment, a Simulation Smoother was used
to generate estimates for the raw speech samples. This
Simulation Smoother included a Kalman filter stage and a
smoothing filter stage in order to generate the estimates
of the raw speech samples. In an alternative embodiment,
20 the smoothing filter stage may be omitted, since the
Kalman filter stage generates estimates of the raw speech
(see equation (33)). However, these raw speech samples
were ignored, since the speech samples generated by the
smoothing filter are considered to be more accurate and
25 robust. This is because the Kalman filter essentially

generates a point estimate of the speech samples from the joint probability density function for the raw speech, whereas the Simulation Smoother draws a sample from this probability density function.

5

In the above embodiment, a Simulation Smoother was used in order to generate estimates of the raw speech samples. It is possible to avoid having to estimate the raw speech samples by treating them as "nuisance parameters" and integrating them out of equation (19). However, this is not preferred, since the resulting integral will have a much more complex form than the Gaussian and Inverse Gamma mixture defined in equation (19). This in turn will result in more complex conditional probabilities corresponding to equations (20) to (30). In a similar way, the other nuisance parameters (such as the coefficient variances or any of the Inverse Gamma, alpha and beta parameters) may be integrated out as well. However, again this is not preferred, since it increases the complexity of the density function to be sampled using the Gibbs sampler. The technique of integrating out nuisance parameters is well known in the field of statistical analysis and will not be described further here.

10

15

20

25

In the above embodiment, the data analysis unit analysed the samples drawn by the Gibbs sampler by determining a histogram for each of the model parameters and then determining the value of the model parameter using a weighted average of the samples drawn by the Gibbs sampler with the weighting being dependent upon the number of samples in the corresponding bin. In an alternative embodiment, the value of the model parameter may be determined from the histogram as being the value of the model parameter having the highest count. Alternatively, a predetermined curve (such as a bell curve) could be fitted to the histogram in order to identify the maximum which best fits the histogram.

In the above embodiment, the statistical analysis unit modelled the underlying speech production process with separate speech source models (AR filters) and channel models. Whilst this is the preferred model structure, the underlying speech production process may be modelled without the channel models. In this case, there is no need to estimate the values of the raw speech samples using a Kalman filter or the like, although this can still be done. However, such a model of the underlying speech production process is not preferred, since the speech model will inevitably represent aspects of the

channel as well as the speech. Further, although the statistical analysis unit described above ran a model order selection routine in order to allow the model orders of the AR filter model and the channel model to vary, this is not essential. In particular, the model order of the AR filter model and the channel model may be fixed in advance, although this is not preferred since it will inevitably introduce errors into the representation.

In the above embodiments, the speech that was processed was received from a user via a microphone. As those skilled in the art will appreciate, the speech may be received from a telephone line or may have been stored on a recording medium. In this case, the channel models will compensate for this so that the AR filter coefficients representative of the actual speech that has been spoken should not be significantly affected.

In the above embodiments, the speech generation process was modelled as an auto-regressive (AR) process and the channel was modelled as a moving average (MA) process. As those skilled in the art will appreciate, other signal models may be used. However, these models are preferred because it has been found that they suitably represent the speech source and the channel they are intended to

model.

In the above embodiments, during the running of the model order selection routine, a new model order was proposed by drawing a random variable from a predetermined Laplacian distribution function. As those skilled in the art will appreciate, other techniques may be used. For example the new model order may be proposed in a deterministic way (ie under predetermined rules), provided that the model order space is sufficiently sampled.